Acoustic Analysis of Czech Expressive Recordings from a Single Speaker in Terms of Various Communicative Functions

Martin Grůber
University of West Bohemia, Faculty of Applied Sciences, Department of Cybernetics
Univerzitní 8, 30614, Plzeň, Czech Republic
gruber@kky.zcu.cz

***Abstract*—This paper presents an extensive acoustic analysis of utterances which were recorded by a single Czech female speaker using various expressive speaking styles. The recording of the expressive utterances was performed as a dialogue between a human and a computer on a given topic. Speech of the human speaker was captured and later carefully transcribed by human annotators. It was also annotated using a listening test. The aim of the annotations was to label each utterance with a corresponding speaking style (referred to as a communicative function). Based on such a labeling, the expressive recordings were classified into various groups and acoustically analyzed. In particular, we placed emphasis on some features which are supposed to influence the perception of speech, such as F0, phoneme duration, formant frequencies or energy. We made an effort to reveal some acoustic differences between the various speaking styles that could help us to improve expressive speech synthesis in a given limited domain.**

I. Introduction

Current speech synthesis techniques produce high quality and intelligible speech. However, the synthetic speech cannot sound completely natural until it expresses a speaker's attitude. Thus, expressive (or emotional) speech synthesis is a frequently discussed topic and has become a concern of many scientists. Even though some results have already been presented, this task has not been satisfactorily solved yet. Some papers which deal with this problem include, but are not limited to [1] [2] [3] or [4].

Since general expressive speech synthesis is a difficult-to-solve task, a limited domain was defined first. Our field of interest was conversation between a human and a computer over family photos (our work is a part of Companions project, whose aim is to develop a virtual companion for conversations). Having the specific limited domain defined, the task of expressive speech synthesis becomes more easily solvable.

To incorporate some expressivity into the synthetic speech, we firstly need to find out which factors are important for listeners to perceive spoken speech as expressive speech. In the first phase of this research, we have focused on acoustic characteristics of the expressive speech. The results of our analysis cannot be generalized as we have analyzed sentences uttered by a single speaker and due to this reason they are not statistically representative. Nevertheless, we can still consider the revealed acoustic characteristics for incorporation of expressivity into our speech synthesis system since it is corpus oriented and based on a unit selection method [5].

For the speech synthesizers employing the unit selection methods, an extensive speech corpus has to be recorded by a single speaker, usually in a consistent speaking style. In short, from the transcribed, annotated and segmented speech corpus (the process of corpus creation is described in [6] or [7]), so-called speech units are taken in such an optimal way that their sequence forms as fluent and natural speech as possible. To get the optimal sequence, two evaluation costs are used. The first one, a target cost, determines a measure of appropriateness of a particular speech unit to a particular location in a synthesized utterance. The second one, concatenation cost, determines smoothness of a concatenation of two consecutive speech units. These two costs together forms an overall cost which is crucial for optimal sequence finding.

It is obvious that for a development of a single synthetic voice for such a speech synthesis system, a speech corpus recorded by a single speaker is required. Thus, an acoustic analysis of speech by a single speaker should be sufficient enough for our purposes.

This paper is organized as follows. Section II deals with a description of data used in the analysis and a description of various expressive styles. In section III, the acoustic analysis of the data is described. In that section, the features that were measured on the data are listed and the techniques which were used for their acquisition are described. In Section IV, some suggestions for further development of the limited domain expressive speech synthesizer are presented. Section V is dedicated to an overview of the attained results. Finally, some future work is presented in Section VI.

II. Description of data to be analyzed

In this section, the description of the data to be analyzed and the various expressive speaking styles - communicative functions - is presented.

A. Expressive data

To become acquainted with the limited domain we are talking about, natural dialogues between humans and a computer were recorded using Wizard-of-Oz method. The process of natural dialogues acquiring is presented in [8]. However, for a better idea, it will be also shortly described here.

In a room that was adapted to the recording purposes, a computer with only a screen and speakers was installed. No other input devices were presented. The human object was equipped by a wireless microphone, another one was used for capturing the computer sound output. This should evoke a feeling that the computer is using an artificial intelligence to communicate with the subject. However, the computer was remotely controlled via a web interface by human operators that were "hidden behind the curtain". Thus, the subject was convinced that he was communicating with the computer just with his voice. During the recording session, subject's family photos were presented by a virtual avatar [9] on the screen and a dialogue between the subject and the avatar (controlled by human operators in fact) was being developed. This way, 65 one–hour sessions were recorded.

Based on these natural dialogues, an expressive speech corpus to be used for unit selection speech synthesis was designed. Actually, it consists of phrases that were originally uttered by the avatar. Later, it was recorded by a professional female speaker - a stage–player - and the process of recording is in details described in [8]. Mainly, the speaker was instructed to put herself in a role of a partner in a dialogue. In order to facilitate such an empathy, a special software application was developed - it played back the parts of the WoZ natural dialogues where the subject was speaking (to provide the speaker with the relevant context) and at the time where the avatar has originally spoken, the dialogue was paused and the speaker was prompted to record the avatar's sentence herself. The text of the actual sentence was displayed on the screen even when the real (context) dialogue was being played so that the speaker had enough time to get acquainted with it before the recording.

Data from this expressive speech corpus were used for the analysis presented in this work.

B. Communicative functions

Communicative functions are supposed to describe various categories of expressivity which can appear in spoken utterances. The set of communicative functions was partially inspired by speech acts presented in [10]. However, the original set was modified, partly extended and adapted to our task during the natural dialogues analysis. The current form of the set of communicative functions is shown in Table I. This set is possibly not the final version. It can still be modified in the future since the characteristics of various communicative functions can be evaluated as equal or very similar in this acoustic analysis and therefore they might be joint together.

For easier orientation, the communicative function names are shortened to understandable labels in the further text.

TABLE I
SET OF COMMUNICATIVE FUNCTIONS

communicative function	*example*
directive	Tell me that. Talk.
request	Let's get back to that later.
wait	Wait a minute. Just a moment.
apology	I'm sorry. Excuse me.
greeting	Hello. Good morning.
goodbye	Goodbye. See you later.
thanks	Thank you. Thanks.
surprise	Do you really have 10 siblings?
sad empathy	I'm sorry to hear that. It's really terrible.
happy empathy	It's nice. Great. It had to be wonderful.
showing interest	Can you tell me more about it?
confirmation	Yes. Yeah. I see. Well. Hmm.
disconfirmation	No. I don't understand.
encouragement	Well. For example? And what about you?
not specified	Do you hear me well? My name is Paul.

III. ACOUSTIC ANALYSIS

There are two main objectives for performing this kind of speech acoustic analysis.

The first aim is to find a correlation between acoustic parameters of expressive speech and the way the speech is perceived by listeners. Annotations of the expressive speech have already been performed [11] and the result is that we have expressive utterances objectively labeled with various communicative functions (speaking styles, further referred to as CFs, more details in Section II-B). Since these annotations show us how the expressive speech is perceived by listeners, the task of this work is to measure the acoustic parameters and to briefly analyze the results.

The second aim is to reveal similarities among various CFs in terms of acoustic parameters that can be used during the process of unit selection speech synthesis (as it was described above) when computing the target cost. On the basis of the measured similarities and dissimilarities, we can create a penalty matrix in the form of

$$\begin{pmatrix} 0.0 & a_{12} & a_{13} & ... \\ a_{21} & 0.0 & a_{23} & ... \\ a_{31} & a_{32} & 0.0 & ... \\ ... & ... & ... & \end{pmatrix}$$

where $a_{ij} = a_{ji}$ represents a penalty (one component of the target cost) for a speech unit marked with CF *i* while required *j* and vice versa (the expressive annotations should be taken into consideration when evaluating specific matrix coefficients either, see Section IV).

This way, more suitable unit can be chosen from the speech corpus in case there is no appropriate unit labeled with the required CF or there is a unit with other CF that suits the optimal sequence better (regarding the other components of the target cost or the concatenation cost). Using a unit of other CF than required during the units concatenation process will be penalized according to the penalty matrix. The design of a particular penalty matrix is one of the future tasks.

In the following sections, results of measurements carried out on all phonemes the particular measurement makes sense for are presented. Since the utterances of various CFs are not phoneme balanced (e.g. utterances of CF *A* might contain more vowels whereas utterances of CF *B* might contain more consonants) and this imbalance might affect the analysis outcome,

the results for phoneme *a* (as an example) are also presented. After all the analyses, correlation coefficients between these two groups of phonemes was calculated, separately for each analysis (except the formant analysis which makes sense only for vowels and each vowel has specific ranges which the formant frequencies vary in). The results in all cases indicate that the groups are (highly) correlated which means that although the CFs are phonemes unbalanced, the values acquired for all phonemes can still be viewed as presentable.

In Table II, numbers of phonemes that were analyzed in various phoneme groups are shown. As it was mentioned above, these total numbers were unbalancedly spread among various CFs. Since the most of the expressive sentences were labeled as *ENCURAGE* or *SHOW-INTEREST*, the most phonemes (approximatelly 70%) fall into these CFs.

TABLE II
NUMBER OF PHONEMES (SEGMENTS) THAT WERE ANALYZED IN VARIOUS PHONEME GROUPS.

phoneme group	approximate number of analyzed phonemes
a phoneme	$10,700$
all phonemes	$134,000$
voiced phonemes	$96,500$

A. F0 analysis

Since in addition to the speech signal the glottal signal was captured too when recording the expressive utterances, we employed a Robust Multi-Phase Pitch-Mark Detection Algorithm [12] to detect pitch pulses in the signal precisely. The F0 values were then derived using this pitch marks sequence. First, we obtained local F0 estimates calculated as median of inverse values of distances between four consecutive pitch marks. Then, the sequence of these local F0 estimates was smoothed by median filter of order 3. Thus, for each phoneme a single mean value of its F0 was determined. The results for a phoneme *a* and separately a mean value for all Czech voiced phonemes for various CFs are shown in Table III.

The results summarized in Table III show that differences between various communicative functions were found. To verify that the differences are statistically significant, a one-way analysis of variance (ANOVA) was performed. The graphical output of the overall analysis is depicted in Fig. 1 (for phoneme *a*). The resulting *p-value* of the analysis was quantified as zero which means the differences are really significant.

The results might suggest that between several CFs (*WAIT*, *APOLOGY*, *SAD-EMPATHY*) for *a* phoneme the differences are not so substantial - the ANOVA *p-value* for these CFs was 0.998. In terms of F0 value, this might suggest possible grouping of these CFs. However, the voiced phonemes is not the same case and also the following results will not confirm this hypothesis.

TABLE III
F0 MEAN VALUES AND STANDARD DEVIATIONS OF PHONEME *a* AND ALL VOICED PHONEMES FOR VARIOUS COMMUNICATIVE FUNCTIONS.

communicative function	*a*		all voiced phonemes	
	mean value [Hz]	*std. dev. [Hz]*	*mean value [Hz]*	*std. dev. [Hz]*
directive	198	21	195	31
request	204	15	204	21
wait	187	4	188	25
apology	187	16	196	16
greeting	184	36	189	25
goodbye	194	21	196	21
thanks	181	13	191	18
surprise	199	23	197	25
sad empathy	187	19	187	19
happy empathy	194	20	187	20
showing interest	203	26	204	29
confirmation	191	34	192	27
disconfirmation	173	32	184	22
encouragement	205	21	205	24
not specified	197	22	196	24

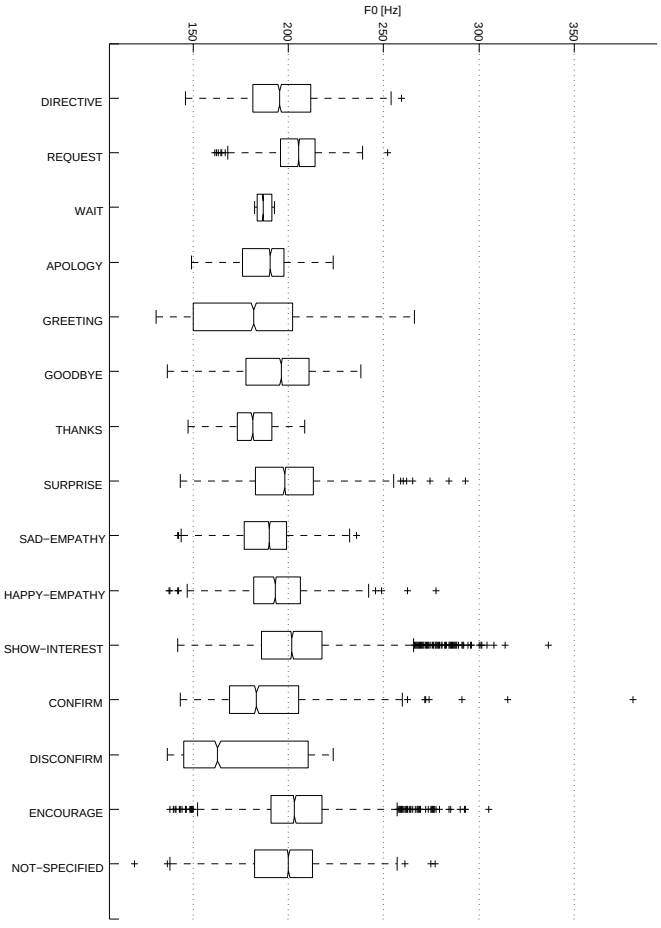

Fig. 1. Results of one-way ANOVA for F0 values of phoneme *a*.

B. Duration analysis

To assign a particular phoneme with a time boundaries clearly, an automatic segmentation technique using HTK Tools improved by a statistic approach [7] was employed. After-

wards, to compute the duration of the phoneme, the times of its end and its beginning were simply subtracted. The mean values for phoneme *a* and for all Czech phonemes are presented in Table IV.

TABLE IV
DURATION MEAN VALUES AND STANDARD DEVIATIONS OF PHONEME *a* AND ALL PHONEMES FOR VARIOUS COMMUNICATIVE FUNCTIONS

communicative function	*a*		all phonemes	
	mean value[ms]	*std. dev.[ms]*	*mean value[ms]*	*std dev.[ms]*
directive	66.5	23.0	90.6	57.3
request	64.9	18.8	88.8	49.8
wait	52.0	7.5	75.4	40.0
apology	100.9	42.3	89.8	52.5
greeting	111.0	52.0	88.4	51.6
goodbye	88.9	34.3	83.7	49.9
thanks	105.4	37.7	99.6	48.8
surprise	86.3	41.2	93.0	50.4
sad empathy	92.7	47.9	95.6	58.8
happy empathy	88.1	38.9	94.7	54.5
showing interest	80.2	39.7	91.8	44.8
confirmation	117.8	54.4	116.3	73.0
disconfirmation	121.2	55.5	109.1	86.3
encouragement	81.1	40.6	90.7	45.5
not specified	85.1	34.5	88.0	50.7

It is obvious that there are remarkable differences regarding the phoneme duration when considering various communicative functions. To prove that the differences between mean values are statistically significant, we decided to perform a statistical one-way ANOVA again. A resulting *p-value* was very close to zero, less than 0.05, which means that the differences can be marked as significant. Also the *p-value* for CFs previously suggested to be grouped (after F0 analysis) was less than 0.05 threshold and thus denying the grouping hypothesis.

Taking into account the mean value of 85.1 ms for CF *NOT-SPECIFIED* (for phoneme *a* measurement), that can be viewed as neutral speech, we can identify communicative functions that contain considerably longer/shorter phonemes meaning slower/faster speech rate.

C. Formant analysis

Formants can be defined as the spectral peaks of the sound spectrum of the voice. Formant is also used to mean an acoustic resonance and a resonance of the human vocal tract. There are many applications that are able to provide formant frequencies estimates from speech signal. We decided to use Speech Filing System[1] as one of its part is `formanal` program which is referred to as currently the best one in SFS to perform fixed-frame formant analysis. This program was originally implemented in the Entropic Signal Processing System and it is used under license from Microsoft.

To each speech segment (a vowel), a single value for formant frequency was assigned. This value came from the middle part of the vowel which was found by cutting the initial and the final quarter of the vowel length. Mean values of the first three formant frequencies of phoneme *a* for various communicative functions are presented in Table V.

TABLE V
THE FIRST THREE FORMANT FREQUENCIES OF PHONEME *a* FOR VARIOUS COMMUNICATIVE FUNCTIONS.

communicative function	*F1 mean value[Hz]*	*F2 mean value [Hz]*	*F3 mean value [Hz]*
directive	635	1581	2798
request	646	1603	2868
wait	659	1695	2508
apology	680	1410	2656
greeting	605	1423	2576
goodbye	671	1602	2816
thanks	649	1467	2652
surprise	642	1456	2685
sad empathy	589	1419	2593
happy empathy	648	1531	2825
showing interest	633	1503	2731
confirmation	627	1403	2492
disconfirmation	658	1467	2873
encouragement	670	1485	2755
not specified	642	1539	2789

In the results, we can identify that the formant frequencies differ between various CFs. Also in this case, one-way ANOVA was performed. The differences between mean values were rated as statistically significant, since the *p-value* was very close to zero. However, there are shown only results for one of the Czech vowels that appear in speech. Considering measurements for all the vowels, the discussion of the results of the formant analysis seems to be too complex and it is out of scope of this paper.

D. Energy analysis

RMS[2] energy is a value that characterizes the intensity of a speech signal. Differences of intensity level between various CFs can be measured using this feature.

For the calculation of the RMS energy of a particular phoneme (speech segment), the Equation 1 was used.

$$RMS = \sqrt{\frac{\sum_{i=1}^{n} s(i)^2}{n}}, \tag{1}$$

where $s(i)$ is $i-th$ sample of the signal and n is the length of the signal.

The results obtained for the expressive recordings are shown in Table. VI. Again, results for phoneme *a* and all phonemes are presented separately.

It is obvious that there are differences between various CFs. In spite of the differences are not so substantial, they are statistically significant which was proved by one-way ANOVA (*p-value* $= 0.0$).

[1]Speech Filing System - http://www.phon.ucl.ac.uk/resource/sfs

[2]RMS = Root Mean Square, also known as the quadratic mean; a statistical measure of the magnitude of a varying quantity. It is especially useful when variates are positive and negative, e.g. waves.

TABLE VI
MEAN VALUES AND STANDARD DEVIATIONS OF THE RMS ENERGY (SIGNAL RANGE $\langle -1,\ 1 \rangle$) OF PHONEME *a* AND ALL PHONEMES FOR VARIOUS COMMUNICATIVE FUNCTIONS.

communicative function	*a*		all phonemes	
	mean value	*std. dev.*	*mean value*	*std. dev.*
directive	0.25	0.06	0.17	0.11
request	0.27	0.05	0.19	0.10
wait	0.19	0.01	0.14	0.09
apology	0.21	0.06	0.19	0.10
greeting	0.16	0.05	0.15	0.09
goodbye	0.25	0.08	0.18	0.10
thanks	0.21	0.06	0.19	0.09
surprise	0.26	0.07	0.19	0.11
sad empathy	0.23	0.08	0.19	0.11
happy empathy	0.25	0.07	0.17	0.10
showing interest	0.25	0.07	0.18	0.11
confirmation	0.18	0.08	0.18	0.12
disconfirmation	0.17	0.10	0.18	0.10
encouragement	0.24	0.07	0.18	0.11
not specified	0.24	0.07	0.17	0.10

IV. SUGGESTIONS FOR EXPRESSIVE SPEECH SYNTHESIS

The achieved results confirmed that all the features measured on the speech signal are important acoustic correlates of various speaking styles. Nevertheless, it cannot be concluded that only these features are sufficient enough for the distinction. In addition, to find acoustic similarities/dissimilarities among various CFs, some extensive statistical tools should be employed yet (e.g. multivariete analysis of variance - MANOVA).

The design of the specific penalty matrix, mentioned in Section III in general, is one of the main tasks for further research. To determine the matrix coefficients, an analysis of expressive annotations of the recordings and finding correlation between the acoustic measures and these annotations is going to be performed. Afterwards, a measure of expressiveness similarity should be designed.

The unit selection algorithm should also undergo a revision of weighs used for various components forming the target cost. The weigh of the component determining the measure of expressiveness similarity among CFs (taken from the penalty matrix) should be set properly taking into consideration all the other components (e.g. using some grid-search methods).

Using this penalty matrix in the right way, the unit selection algorithms should be able to generate the optimal speech unit sequence more precisely. Thus, the expressivity should be perceived in the synthetic speech while keeping naturalness and intelligibility at acceptable level. The incorporation of such a matrix and modifications in the unit selection algorithms regarding this issue are considered as our future work.

V. SUMMARY

In this work, we acoustically analyzed speech units coming from the expressive speech corpus that was recorded by a single Czech female speaker under specific circumstances. The important acoustic measures as F0, phoneme duration, formant frequencies and RMS energy were observed.

The speech corpus is planned to be used for expressive speech synthesis in the limited domain of conversations between a human and a computer on the given topic. According to this particular domain, a set of communicative functions (speaking styles) was defined. We tried to reveal acoustic differences between utterances labeled by various communicative functions and to suggest some enhancements for unit selection algorithms. These suggestions should head towards expressive speech synthesis and may improve naturalness of such synthetic speech.

Another interesting output was detected regarding the relationship among features we measured on the speech signal of various CFs. In Table VII and Table VIII, correlation coefficients calculated among the various features are presented. For the calculation, normalized values were used.

TABLE VII
CORRELATION COEFFICIENTS AMONG VARIOUS FEATURES MEASURED FOR ALL PHONEMES.

feature	F0	duration	RMS
F0	1.00	-0.30	0.34
duration	-0.30	1.00	0.45
RMS	0.34	0.45	1.00

TABLE VIII
CORRELATION COEFFICIENTS AMONG VARIOUS FEATURES MEASURED FOR PHONEME *a* (*F1* STANDS FOR THE FIRST FORMANT FREQUENCY, ETC.).

feature	F0	duration	F1	F2	F3	RMS
F0	1.00	**-0.62**	0.08	0.30	0.29	**0.80**
duration	**-0.62**	1.00	-0.16	**-0.81**	-0.15	**-0.60**
F1	0.08	-0.16	1.00	0.35	0.38	0.15
F2	0.30	**-0.81**	0.35	1.00	0.30	0.35
F3	0.29	-0.15	0.38	0.30	1.00	0.55
RMS	**0.80**	**-0.60**	0.15	0.35	0.55	1.00

The results show that for all phonemes, there was not found any high correlation, slight one can be identified between duration and RMS energy features. For phoneme *a*, the situation is only a little bit different. The correlations can be described as stronger but keeping the trend. The only exception is the relationship between duration and RMS energy that is inverse. Although the correlation coefficients can only detect linear dependencies among the variables, these results are worth further research.

VI. FUTURE WORK

Apart from the suggestions for the further development of the expressive speech synthesis system already mentioned in Section IV as the penalty matrix creation or the target cost weighs revision, there are other interesting issues resulting from the expressivity research.

One of the tasks for further development should be to try to extend the set of communicative functions to cover wider range of conversational topics. Ideally, the set would be well-formed enough to cover naturally spoken speech in general.

In addition, the results of this analysis might be used for determining communicative function from speech signal. The

values of F0, duration, formant frequencies or RMS energy can be used as predictors for such a classification task. In the case that any classification model were able to give good results using these predictors, we could conclude that it would be sufficient to model these characteristic of speech to obtain expressive output (applicable e.g. for HMM speech synthesis systems [13]).

ACKNOWLEDGEMENTS

This research was funded by the Grant Agency of the Czech Republic, project No. GAČR 102/09/0989 and partly also by the University of West Bohemia, project No. SGS-2010-054. The access to the METACentrum computing facilities provided under the research intent MSM6383917201 is highly appreciated.

REFERENCES

[1] A. W. Black, "Unit selection and emotional speech," in *Proceedings of Eurospeech*, Geneva, Switzerland, 2003, pp. 1649–1652.

[2] M. Bulut, S. S. Narayanan, and A. K. Syrdal, "Expressive speech synthesis using a concatenative synthesiser," in *Proceedings of the 7th International Conference on Spoken Language Processing – ICSLP*, Denver, CO, USA, 2002, pp. 1265–1268.

[3] W. Hamza, R. Bakis, E. M. Eide, M. A. Picheny, and J. F. Pitrelli, "The IBM expressive speech synthesis system," in *Proceedings of the 8th International Conference on Spoken Language Processing – ISCLP*, Jeju, Korea, 2004, pp. 2577–2580.

[4] I. Steiner, M. Schder, M. Charfuelan, and A. Klepp, "Symbolic vs. acoustics-based style control for expressive unit selection," in *Seventh ISCA Tutorial and Research Workshop on Speech Synthesis*, Kyoto, Japan, September 2010, pp. 114–119.

[5] J. Matoušek, D. Tihelka, and J. Romportl, "Current state of Czech text-to-speech system ARTIC," in *Text, Speech and Dialogue, proceedings of the 9th International Conference TSD 2006*, ser. Lecture Notes in Computer Science, vol. 4188. Berlin, Heidelberg: Springer, 2006, pp. 439–446.

[6] J. Matoušek and J. Romportl, "Recording and annotation of speech corpus for czech unit selection speech synthesis," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin–Heidelberg, Germany: Springer, 2007, vol. 4629, pp. 326–333.

[7] J. Matoušek, D. Tihelka, and J. Psutka, "Experiments with automatic segmentation for czech speech synthesis," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2003, vol. 2807, pp. 287–294.

[8] M. Grůber, M. Legát, P. Ircing, J. Romportl, and J. Psutka, "Czech Senior COMPANION: Wizard of Oz data collection and expressive speech corpus recording and annotation," in *Human Language Technology. Challenges for Computer Science and Linguistics*, ser. Lecture Notes in Computer Science, Z. Vetulani, Ed., vol. 6562. Berlin-Heidelberg, Germany: Springer, 2011, pp. 280–290.

[9] Z. Krňoul and M. Železný, "Realistic face animation for a czech talking head," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, P. Sojka, I. Kopecek, and K. Pala, Eds. Berlin-Heidelberg: Springer, 2004, vol. 3206, pp. 603–610.

[10] A. K. Syrdal and Y.-J. Kim, "Dialog speech acts and prosody: Considerations for TTS," in *Proceedings of Speech Prosody*, Campinas, Brazil, May 2008, pp. 661–665.

[11] M. Grůber and J. Matoušek, "Listening-test-based annotation of communicative functions for expressive speech synthesis," in *Text, Speech and Dialogue, proceedings of the 13th International Conference TSD 2010*, ser. Lecture Notes in Computer Science, vol. 6231. Berlin-Heidelberg, Germany: Springer, 2010, pp. 283–290.

[12] M. Legát, J. Matoušek, and D. Tihelka, "On the detection of pitch marks using a robust multi-phase algorithm," *Speech Communication*, vol. 53, no. 4, pp. 552–566, 2011.

[13] Z. Hanzlíček, "Czech HMM-based speech synthesis," in *Text, Speech and Dialogue, proceedings of the 13th International Conference TSD 2010*, ser. Lecture Notes in Computer Science, vol. 6231. Berlin-Heidelberg, Germany: Springer, 2010, pp. 291–298.